Technical Report

Department of Computer Science
and Engineering
University of Minnesota
4-192 EECS Building
200 Union Street SE
Minneapolis, MN 55455-0159 USA

TR 04-028

Finding Functionally Related Genes by Local and Global Analysis of MEDLINE Abstracts

Sigve Nakken, Christopher Kauffman, and George Karypis

June 29, 2004

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
29 JUN 2004		**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Finding Functionally Related Genes by Local and Global Analysis of MEDLINE Abstracts	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Army Research Laboratory,2800 Powder Mill Road,Adelphi,MD,20783-1197**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**10**	

Finding Functionally Related Genes by Local and Global Analysis of MEDLINE Abstracts

Sigve Nakken[*], Christopher Kauffman, George Karypis

Department of Computer Science and Engineering, University of Minnesota, Twin Cities[†]
Digital Technology Center
499 Walter Library, 117 Pleasent St SE
Minneapolis, MN

nakken@cs.umn.edu, kauffman@cs.umn.edu, karypis@cs.umn.edu

ABSTRACT

Discovery of biological relationships between genes is one of the keys to understanding the complex functional nature of the human genome. Currently, most of the knowledge about interrelating genes are found in immense amounts of various biomedical literature. Hence, extraction of biological contexts occurring in free text represents a valuable tool in gaining knowledge about gene interactions. We present a textual analysis of documents associated with pairs of genes, and describe how this approach can be used to discover and annotate functional relationships among genes. A study on a subset of human genes show that our analysis tool can act as a ranking mechanism for sets of genes based on their functional relatedness.

Keywords

information retrieval, document clustering, gene relations

1. INTRODUCTION

Although most genes in the human DNA now have been completely sequenced [3], their functional roles and the diverse interrelationships between them are still to be fully understood. With the development of the DNA microarray [10], researchers have a tool where they can measure the expression levels of several genes at a time. Producing huge amounts of data, discoveries made from such experiments are published at an enormous rate in the scientific literature; thus, giving researchers a severe information retrieval challenge in keeping up to date in their fields of expertise. With the aim of structuring existing knowledge occurring in free text, biomedical text collections have been subject to extensive research the last years.

The problem of extracting information about how genes are related has been the major focus by many groups, e.g. [5, 8, 12, 16, 17, 18], and has led to a variety of approaches for discovery of functional groupings among genes. Clearly, any successful method should be able to extract the biological nature of the discovered relationships. This goal has been achieved to some extent by different efforts, but they either rely on the quality of documents associated with genes [12] or limit themselves to controlled vocabularies for annotating the relationships [5, 16]. Furthermore, knowledge about gene relations often include several biomedical aspects, i.e. biology, chemistry and medicine. This fact reflects the complex nature of gene relationships, and indicates that they ought to be characterized by more than one functional context.

We propose an approach that initially extracts the multiple local contexts between pairs of genes found co-occurring in MEDLINE abstracts. Further, a global analysis of local contexts between pairs is performed, giving similiar local contexts a global interpretation. It is our belief that this scheme can represent an efficient way of discovering functionally related genes.

We evaulate our method on a subset of human genes, and the results (though preliminary) show that sets of *genes* connected by same global contexts are functionally similar.

The rest of the paper is organized as follows: The next section presents related work on mining the literature for gene-relations. We then give a description of the models and methods used in our scheme for finding functionally related genes. Finally, we present and discuss preliminary results on applying our approach on a set of human genes.

[*]Exchange student from the Norwegian University of Science and Technology (NTNU).

[†]This work was supported in part by NSF ACI-0133464 and ACI-0312828; the Digital Technology Center at the University of Minnesota; and by the Army High Performance Computing Research Center (AHPCRC) under the auspices of the Department of the Army, Army Research Laboratory (ARL) under Cooperative Agreement number DAAD19-01-2-0014. The content of which does not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred. Access to research and computing facilities was provided by the Digital Technology Center and the Minnesota Supercomputing Institute.

2. RELATED WORK

Detecting gene relations based on the co-occurrence methodology was initially explored by Stapley et. al [14] in their prototype system for visualization of gene interactions. Later, the method was utilized in a comprehensive manner by Jenssen et. al [5, 6], who developed a genome-wide network of human genes. The co-occurrence method is very efficient for its purpose; to *detect* gene relations. However, co-occurrence alone can not help us in discovering the *characteristics* of the relation. An approach of going beyond simple co-occurrence was suggested by [5], who annotated the relations between genes detected by co-occurrence with associated MeSH and GO terms.

Recently, analysis of the graph structure inherent in a co-occurrence network has attracted the attention of researchers, e.g. [17, 18]. Wilkinson et al. [17] employed Girvan and Newman's process of finding communities [4] to discover related genes. By picking sets of genes statistically correlated to user-selected keywords, components of a gene co-occurrence graph are partitioned into functionally related communities. Interesting results included placing co-occurring genes into different communities; demonstrating the fact that co-occurrence does not always imply functional relatedness. Wren et. al [18] took advantage of statistical properties of connections in the network to determine the "cohesiveness" of sets of co-occurring objects (genes, diseases, chemical compounds etc.). The technique could therefore identify whether a set of objects form a purposeful grouping, and maybe more importantly, whether members not in the set should be included.

Based on domain knowledge from thesauri, Stephens et. al [16] both found and annotated gene relationships by scanning sentences for gene thesauri terms. However, the approach is dependent upon high quality domain-specific thesauri in order to produce good results.

Given a group of genes, Raychaudhuri et. al [8] developed the concept of *neighbor divergence pr. gene(NDPG)* within scientic texts to discover a potential biological relation in the group. The motivation behind their approach was to recognize articles describing the function inherent in the group. It achieved accurate results on a testset taken from the yeast organism (79% recall at 100% precision). However, the method requires that a list of relevant articles is provided for each gene in the organism, and this requirement is by no means trivial. Furthermore, NDPG does not tell us the function among a set of genes, it merely determines if the group shares one.

Approaches using the published literature as the main source for annotation have been investigated earlier. With the same goal as [5] of establishing functional gene relations on a genome-wide scale, Shatkay et. al [12] employed document similiarity search as basis for their method. Arguing that clustering of co-expressed genes from DNA microarry experiments may fail to give the true picture of interrelationships between genes, they proposed a complementory method in which relationships between genes are found and annotated by measuring the similarity between the genes' set of relevant documents in the literature. The annotation mechanism involves a "theme-based" probabilistic search [13], which provides a summary of the content between a query document and its similar documents. The main limitation of this approach is that it requires each gene to be associated with a *kernel document*, capturing most of the gene's functional biology. The method relies heavily on the quality of these documents, which may be hard to find.

3. METHODS

In this section, the methods and models used in our approach are described in more detail.

3.1 Overview

Our work represents a novel method for annotating the *functional contexts* that exist between genes found co-occurring in MEDLINE records. After creating a co-occurrence graph of human genes from MEDLINE, contexts between genes are assigned by *local and global analysis* of documents associated with the edges of the graph. The documents associated with an edge of the graph are the MEDLINE abstracts where a pair of genes co-occurred. First, documents relating to the gene-pairs are clustered into k local clusters; thus, splitting literature related to a pair into k contexts. Furthermore, each cluster (context) between a genepair is associated with its hundred most *descriptive features*. Viewing this operation within the context of the co-occurrence graph, each edge is being split into a multiedge, reflecting multiple relationships between the connecting nodes. Using our terminology, the co-occurrence graph has been *unfolded*.

With the goal of creating a limited set of contexts between the genes in our unfolded graph, we give each edge in the graph a globally defined context, or "color". The colors are defined on the basis of the total set of local contexts occurring between the genes. More specifically, we cluster the total set of descriptive features into a predefined number of clusters. As in the first stage, each cluster (color) is associated with its most descriptive features. This second stage ensures that similar local functional contexts occurring between any pair of genes are given the same global context.

Having a co-occurrence graph between genes as the only prerequisite, our approach of mining gene relations can henceforth achieve two major goals:

- *annotate multiple relationships between pairs of genes* with globally defined functional contexts
- *find functionally related groups of genes* by means of extracting same-colored edges in the colored unfolded co-occurrence graph

3.2 Creation of co-occurrence graph

A co-occurrence network between human genes forms the backbone of our method. As shown in various experiments [5, 6], the co-occurrence method has proved to be an efficient as well as valid approach of detecting meaningful biological relationships between genes. The methodology is simplistic; if two genes co-occur in an abstract, they are assumed to have a relationship of some kind. Our work is no attempt of copying the comprehensive network developed by the people behind PubGene [5], henceforth, we do not intend to improve upon the method for co-occurrence extraction.

In fact, we only used HGNC[1], HUGO Gene Nomenclature Commmitte, as the database of gene symbols used in our search for co-occurrences. That said, the nomenclature provided by HGNC does include literature aliases for a major part of the symbols, and these were also being searched for. Common abbreviations used in biology literature (i.e. IV, SD, ABO etc.) that coincided with gene symbols led to false positives, as experienced by [5, 17]. The actual extraction process was done in a straigthforward manner; whenever a symbol was found in a MEDLINE record (title or abstract), this was considered a match for the gene associated with the gene. A link was made between a pair of genes if they occurred in the same record, and the strength of the link was found by counting the number of records in which the pair co-occurred.

There is, however, a key difference between our extraction process and the one by Jenssen et. al [5]. Along with creating the co-occurrence-based links, the set of MEDLINE records (hereafter termed *documents*) associated with each pair of genes were stored for further analysis.

We model the documents in our collection using the document vector model [2]. This model considers a document as a set of representative keywords, *index terms*. Index terms are document words (mainly nouns) used to summarize the semantic contents of the text. In order to reduce the influence of very common words, the terms are weighted with the TF-IDF (Term Frequency-Inverse Document Freqency) strategy. If M denotes the number of distinct index terms in our collection of N documents, each document i will be represented by a vector on the following format:

$$\vec{d_i} = (w_{i,1}, w_{i,2}, ..., w_{i,M})$$

Each weight $w_{i,j}$ is given by $TF \times IDF$:

$$w_{i,j} = tf_{i,j} \times \log \frac{N}{n_j}$$

where $tf_{i,j}$ is the normalized frequency of term j in document i. The IDF factor is calculated as $\log \frac{N}{n\,j}$, where n_j denotes the number of documents where term j is occurring.

Similarity between two documents are found by seeing how well their two respective vectors correlate, quantified by the cosine of the angle between them:

$$cos(\vec{d_i}, \vec{d_j}) = \frac{\vec{d_i} \bullet \vec{d_j}}{|\vec{d_i}| \times |\vec{d_j}|} \quad (1)$$

The cosine coefficient will range from 0 to 1, where 1 denotes complete similiarity ($\vec{d_i} = \vec{d_j}$), and 0 implies orthogonal vectors.

3.3 Unfolding the co-occurrence graph with document clustering

Along each edge in the co-occurrence graph, we use a clustering software toolkit named CLUTO[2] to cluster the documents into k clusters. At the moment, we use k=2 on every edge, but we are investigating more advanced ways of deciding k (see Section 5). The clustering technique employed is *bisecting K-means*. With the bisecting k-means approach, a document collection is first clustered in two groups, then one of these groups is seleced and bisected further. The similiarity function used for the clustering is the cosine coefficient, given in Equation 1. A detailed explanation of the bisecting K-means clustering technique can be found elsewhere, e.g. [15]. The *most descriptive features* of a cluster is found by selecting the l words the contribute the most to the average similarity between the documents in the cluster. Currently, l=100 is used as the number of descriptive features.

Although we now cluster each edge into k=2 clusters, an extra step is taken to certify that the edge clusters have a certain degree of dissimilarity. If the majority of each cluster's descriptive features are identical, the edge is not clustered into two clusters. This case reflects the fact that all the literature discussing the pair of genes are basically referring to the same context. In order to retrieve such an edge's descriptive features, we treat all its documents as belonging to a single cluster.

3.4 Coloring the unfolded graph

The result of the first clustering stage is a graph with multiple edges between nodes, and where each edge is associated with ten descriptive features. To assign each edge a globally defined color, we cluster the *total set of descriptive features* in the graph into m clusters (colors). The variable m will reflect how many functional contexts we expect to see on a global basis in the graph, and is a factor we are currently experimenting with (see Section 5 for further discussion). As in the first stage, we employ bisecting K-means as the clustering technique. Furthermore, each of the m colors are given descriptive features following the same procedure as in Section 3.3. A color's ten most descriptive features provides a brief summary of a global functional context. Since each edge in the unfolded co-occurrence graph now belongs to particular color and its associated features, the graph has been *colored*.

Given a clique of nodes in the colored unfolded co-occurrence graph, we developed a simple measure of "color purity"; the maximum number of edges in the clique connected by the same color. Since the coloring process can give a gene-pair two global contexts of the same color, two same-colored edges between a gene-pair in a clique were merged into one edge. In that manner, all the k(k-1)/2 gene-pairs in a clique of size k were connected either by two edges of different colors or by one edge alone. A formal expression of the purity measure can then be given:

$$maxColorFrac_c = \frac{argmax_{color} EDGES_c \cdot 2}{k(k-1)}$$

where $EDGES_c$ represents the total set of edges in the unfolded colored clique c of size k.

3.5 GO-similarity

We use the Gene Ontology (GO)[3] as means of validating our method. Being the most comprehensive ontology used

[1] *http://www.gene.ucl.ac.uk/nomenclature/*

[2] *http://www.cs.umn.edu/~karypis/cluto/*

[3] *http://www.geneontology.org*

to describe the functional roles of genes, it is valuable tool for assessing whether two genes are biologically related. The terms comprising GO is organized into a directed acyclic graph (DAG), which has the property of multiple inheritance. Hence, every GO term follows *the true path rule*: if a child term describes a gene product, then all its parents also apply to that gene product. Using EBI's[4] existing GO-annotation of the human genome, we managed to associate 10030 HUGO gene symbols with GO terms. In an effort to expand the number of GO terms pr. gene, we took advantage of the *the true path rule* inherent in the ontology graph structure to generate greater sets of GO terms pr. gene.

One way of measuring gene functional similiarity would be to find which GO terms are common between the genes in question. While this approach is simple and intuitive, clearly it doesn't give us any quantiative measure of similarity. Alternatively, one can consider each gene as a "document", where the document consists of textual descriptions of GO terms associated with it. Furthermore, we can model each document in the vector-space of GO terms, and as shown earlier, this view gives us an opportunity to compute quantitative similarities. Now, the index terms consists of all GO terms associated with gene symbols. If there are a total of N GO terms used in annotation of our genes, we can represent a gene with the following GO vector, where $w_{i,j}$ is the weight of GO term j for gene i:

$$\vec{g_i} = (w_{i,1}, w_{i,2}, ..., w_{i,N})$$

The weighting strategy becomes slightly different than in the ordinary text document context. Since no GO term is associated with a gene more than once, the TF (term frequency) factor is omitted. Thus, each $w_{i,j}$ will only contain the IDF-part, reflecting how relevant or specific GO term j is to gene i. Finally, each gene's weighted GO vector $\vec{g_i}$ is normalized to a vector of length 1. Using the cosine coefficient used previously for abstracts, we can define our notion of GO similiarity.

Definition 1. Given two genes i and j, represented by their *weighted normalized* GO vectors $\vec{g_i}$ and $\vec{g_j}$, their *GO-similarity score* is given by:

$$GOsim(\vec{g_i}, \vec{g_j}) = \vec{g_i} \bullet \vec{g_j}$$

Definition 2. Given a set of n genes, represented by their *weighted normalized* GO vectors ($\vec{g_1}$ ... $\vec{g_n}$), the *average pairwise GO-similarity* in the set is given by the standard sum-of-pairs score:

$$avgGOsim(\vec{g_1}..\vec{g_n}) = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{i>j}^{n} GOsim(\vec{g_i}, \vec{g_j}) \quad (2)$$

We believe this definition of GO-similiarity is valid for our limited purposes. Lord et. al [7] used a similar line of attack when they explored semantic similiarities across GO using Resnik's [9] notion of *shared information content*. Their method was validated by showing that semantic GO similiarity correlated well with sequence similarity in the SWISS-PROT database.

[4] *ftp://ftp.ebi.ac.uk/pub/databases/GO/goa*

4. RESULTS

Our initial co-occurrence graph contained 5799 human genes connected by 73729 edges, each associated with two or more documents. In order to make sure that gene-pairs were represented appropriately in the literature, we pruned the graph to only include edges with between 10 and 100 documents. Furthermore, we kept only genes that were annotated with GO terms, reducing the graph even more. Finally, after removing genes that were considered to be false positives because of bad aliases, our testgraph contained 1516 genes, connected by 4849 edges.

Figure 1 shows a part of the unfolded co-occurrence graph, each edge being denoted with its descriptive features. Note that one connection in this part of the graph contains only one cluster, this corresponds to the case where the documents are considerered to contain only one functional context. However, this illustrates a rare case among gene-pairs in the graph, since almost every pair were given two local contexts during the first clustering process.

After coloring the clusters of edges with the method outlined in Section 3.4 (using m=100 colors), the clique shown in Figure 1 turned into the clique shown in Figure 2. As can be seen from Figure 2, some of the color descriptions are fairly similar, and this observation may imply that a color-scheme with lower number of colors should have been used (see Section 5 for further discussion).

To validate our method on a large scale, we looked at all the *cliques* in the colored unfolded co-occurrence graph. Cliques are fully connected components of the graph, and based on the co-occurrence assumption, a clique can potentially contain a set of functionally related genes. By measuring the color distribution among edges in cliques of size 4 and greater, we investigated whether this distribution were related to functional similarity among the genes in the clique. More specifically, the color purity measure developed in Section 3.4 were used to give rankings among sets of genes. This were accomplished by sorting all the cliques in the graph based on decreasing order of *maxColorFrac*, and plotting the running average GO-similarity of cliques in this ordering.

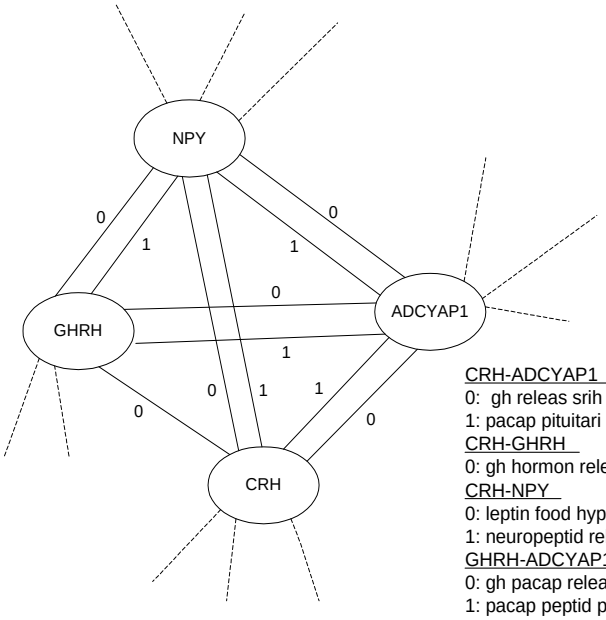
To evaluate the quality of our approach, we compared our color-based ranking with three schemes that only employ local contexts to evaluate a group of genes' relatedness. The first scheme computes *average pairwise document similarity between documents supporting each gene-pair* in a clique of genes:

$$SIM_A = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{i>j}^{n} \frac{D_i \cap D_j}{D_i \cup D_j},$$

where D_i is the set of documents supporting edge i, and n represents the number of edges in the clique. The second one measures *average pairwise textual similarity between documents supporting each gene-pair* in a clique of genes:

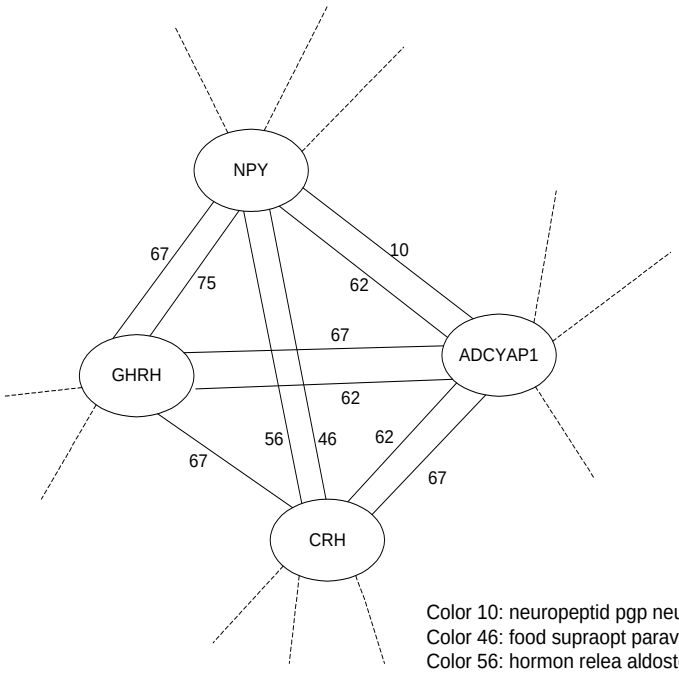
$$SIM_B = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{i>j}^{n} cos(\vec{md_i}, \vec{md_j}),$$

where n denotes the number of edges in the clique, and md_i represents edge i's *metadocument*, a combined document of the documents supporting edge i. The last scheme computes



CRH-ADCYAP1
0: gh releas srih somatotrop pacap38 corticotrop hormon turbot teleost pituitari
1: pacap pituitari peptid cyclase-activ polypeptid adenyl hormon effect placenta rat
CRH-GHRH
0: gh hormon releas sleep srih hypothalam cortisol pituitari microgram secret
CRH-NPY
0: leptin food hypothalam intak neuropeptid energi feed nucleu obes orexigen
1: neuropeptid releas dsip gh cortisol suicid hormon plasma antidepress acth
GHRH-ADCYAP1
0: gh pacap releas srih somatotrop secret pituitari miapaca-2 antagonist hormon
1: pacap peptid pituitari fish cdna polypeptid goldfish adenyl gh receptor
GHRH-NPY
0: gh neuron srih ghs-r releas hypothalam hormon arcuat pituitari mrna
1: iugr intak food punish growth ep peptid retard fr unpunish
NPY-ADCYAP1
0 pacap nerv fiber fibr peptid neuropeptid pacap-27 pacap-38 ganglia sp
1 pacap pacap38 neuron peptid pituitari gpalpha releas adenyl neuropeptid sympathet

Figure 1: A clique of size=4 in the unfolded co-occurrence graph showing genes NPY, GHRH, ADCYAP1 and CRH. Each edge's documents have been clustered into k=2 or k=1 clusters. Also shown is the most descriptive features (stemmed) of each cluster in the clique.



Color 10: neuropeptid pgp neuron nerv innerv som fiber fibr immunoreact cgrp
Color 46: food supraopt paraventricular arcuat fo mrna nucleu pvn neuron hypothalam
Color 56: hormon relea aldosteron plasma pituitari secret acth cortisol adren microgram
Color 62: stimul secret vasoact polypeptid hormon inhibin pituitari peptid pacap adenyl
Color 67: pituitari peptid gnrh srih prolactin somatotrop relea hormon secret gastrin
Color 75: cow supplement oil rumin steer forag diet intak digest fed

Figure 2: The same clique as in Figure 1, now each cluster has been replaced with a colorlabel, the global functional context. The descriptive features (stemmed) of each color is also given.

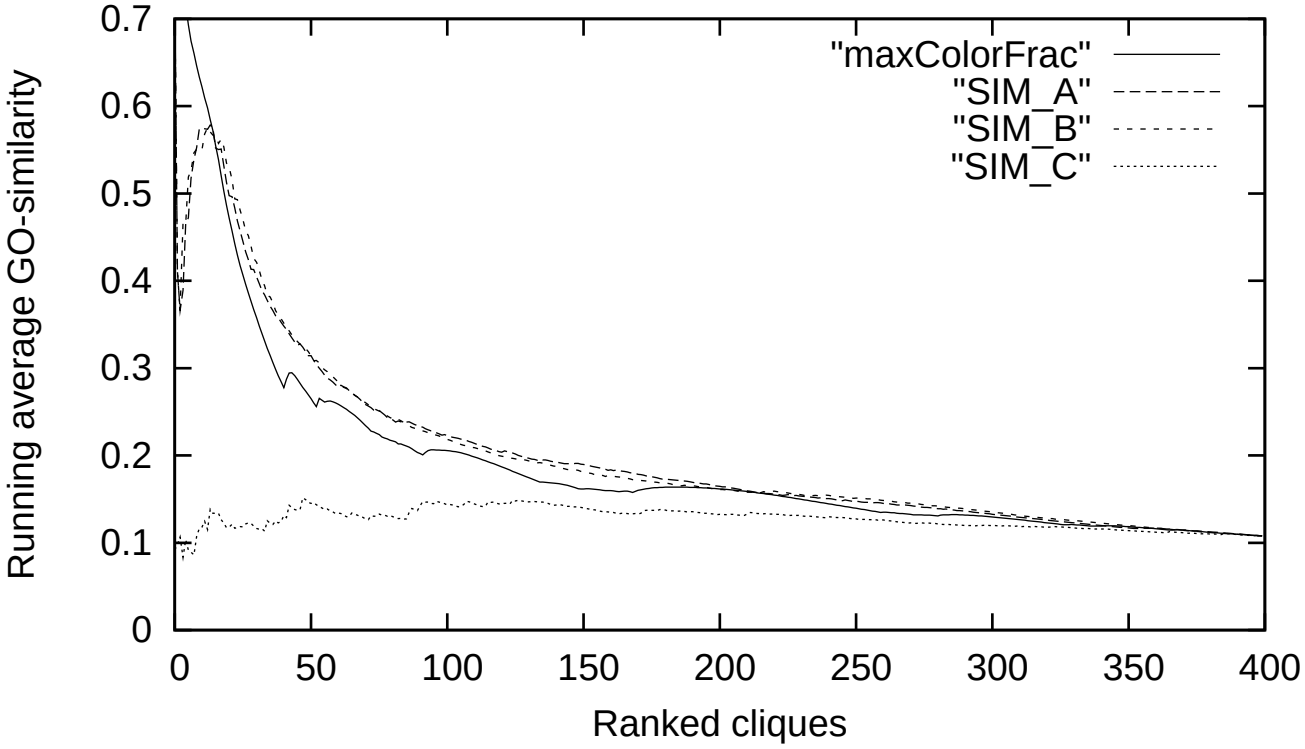


Figure 3: Running average GO-similarity in four different rankings of cliques. The rankings are based on four different methods for determining functional relatedness of the genes in a clique: maxColorFrac = Maximum fraction of clique-edges covered by the same color, SIM_A = Average pairwise document similarity among the edges in the clique, SIM_B = Average pairwise textual similarity among between documents supporting the edges in the clique, and SIM_C = Textual similarity of the union of documents among edges in a clique.

Cliquesize	*Multicolor fraction*
4	56.5%
5	66.7%
6	79.9%
7	85.5%

Table 1: Average fraction of multicolored edges in cliques.

the *average textual similarity of the union of documents supporting each gene-pair* in a clique of genes:

$$SIM_C = \frac{2}{k(k-1)} \sum_{i=1}^{k} \sum_{i>j}^{k} cos(\vec{d_i}, \vec{d_j}),$$

where k is the number of unique documents in the clique, and d_i represents document i in this unique set.

As can be seen from Figure 3, on 5% of the cliques ranked best by the different methods, our scheme discovers more functionally related sets of genes than the other methods. However, on the remaining cliques, the performance of our scheme does not persist in the same manner, and the reason for this is currently being investigated.

To give an indication of how multicolored our cliques are, Table 1 shows the average fraction of multicolored edges in cliques of different sizes. Considering the fact that nearly every gene-pair in our unfolded co-occurrence graph were assigned two edges (local contexts), the global coloring has made sure that similar local contexts are given same global contexts; representing the samecolored fraction of edges. So, even though the majority of genes in cliques are connected with different global contexts, our approach can still find the cliques with the most functionally related genes.

5. DISCUSSION AND FURTHER WORK

There are several limitations to our approach, and it is currently being explored in different ways. Document clustering represents a high-level method for the problem of finding functional contexts between genes, as it does not involve any form for advanced NLP processing. Thus, results should give perspective rather than detailed knowledge. The descriptive features associated with the global contexts exemplifies this in not being very detailed.

The number of local contexts that are likely to exist between a pair of genes will be dependent upon how much research and published literature there is about the pair, and this varies widely for different pairs. Providing each edge with an estimate of k, the number of local contexts likely to exist between the connecting genes, will give the clustering process a higher degree of validity. Intuitively, the number of MEDLINE records between a pair will give some indication for k. Using the MeSH[5] terms associated with each MEDLINE article may also be of importance. Sehgal et. al [11] recently deveveloped MeSH profiles of topics in MEDLINE collections. Developing a MeSH profile for a genepair can act as a heuristic leading to the right size of k.

We will also work on methods for determining the appropriate number of global functional contexts (colors) for a given set of gene pairs. At the moment, we experiment with different colorschemes, and evaluate a scheme's goodness based on empirical observations of the specificity of the different colors' descriptive features. A more theoretical procedure for this assessment would be beneficial for the method's applicability. Factors such as graph size and functional diversity among the genes in the graph will play a significant role in determining the right size of m.

Our results have shown that groups of highly "GO-similar" genes are connected with similar global functional contexts. However, GO-similarity may not give the whole true picture of a set of genes' relatedness with respect to MEDLINE records. Since the literature about gene relations are discussed in a variety of contexts, the functional contexts assigned to a pair of genes will represent a *broad* notion of biomedical knowledge. GO terms, on the other hand, are specific and merely related to genes' biological processes, molecular function and cellular component. Hence, some cliques may appear with high *maxColorFrac* (implying context-related genes) even though their GO-similarity is low.

The mechanism for selecting potential sets of related genes in the graph will influence the functional discoveries among the genes. Although the results by using *cliques* are promising, tracking same-colored connected compontents might give other interesting findings. Moreover, our current measure for the functional relatedness of genes in a clique, *maxColorFrac*, maybe too simple for capturing the properties between the set of genes. A closer investigation of the color distribution in the clique might reveal other functional relations.

As noted earlier, the co-occurrence process has not been our area of focus; thus, our initial graph did possibly include more false positives than desirable. Badly designed gene symbols, coinciding with other abbreviations in the literature, is a matter of great frustration among text miners in biology. Recently, an approach to address and resolve such symbol ambiguities was proposed by Adar [1].

[5] *Every MEDLINE record is associated with Medical Subject Headings (MeSH) terms. See http://www.nlm.nih.gov/mesh/meshhome.html for more information*

6. REFERENCES

[1] E. Adar. Sarad: A simple and robust abbreviation dictionary. *Bioinformatics*, 20(4):527–33., March 2004.

[2] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, New York, NY, USA, 1999.

[3] Genomics and Its Impact on Science and Society: The Human Genome Project and Beyond. http://www.ornl.gov/sci/techresources/Human_Genome, March 2003. Published by U.S. Dept. of Energy.

[4] M. Girvan and M. Newman. Community structure in social and biological networks. In *Proceedings of the National Academy of Sciences*, volume 99, pages 8271–76., 2002.

[5] T. K. Jenssen, A. Laegreid, J. Komorowski, and E. Hovig. A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, 28(1):21–28, 2001.

[6] T. K. Jenssen, L. Öberg, M. L. Anderson, and J. Komorowski. Methods for large-scale mining of networks of human genes. In *Proceedings SIAM Conference Data Mining*, 2001.

[7] P. W. Lord, R. D. Stevens, A. Brass, and C. A. Goble. Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. *Bioinformatics*, 19(10):1275–83., 2003.

[8] S. Raychaudhuri, H. Schutze, and R. Altman. Using text analysis to identify functionally coherent gene groups. *Genome Research*, 12(10):1582–90., 2002.

[9] P. Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95–130., 1999.

[10] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarry. *Science*, 270:467–70., 1995.

[11] A. Sehgal, X. Y. Qiu, and P. Srinivasan. Mining medline metadata to explore genes and their connections. In *Proceedings of SIGIR 2003 Workshop on Text Analysis and Search for Bioinformatics*, 2003.

[12] H. Shatkay, S. Edwards, W. J. Wilbur, and M. Boguski. Genes, themes and microarrays: Using information retrieval for large-scale gene analysis. In *Proc Int Conf Intell Syst Mol Biol*, volume 8, pages 317–28., 2000.

[13] H. Shatkay and W. J. Wilbur. Finding themes in medline documents: Probabilistic similarity search. In *Proceedings of IEEE Advances in Digital Libraries*, pages 183–92., 2000.

[14] B. J. Stapley and G. Benoit. Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in medline abstracts. In *Pacific Symposium on Biocomputing*, volume 5, pages 529–40, 2000.

[15] M. Steinbach, G. Karypis, and V. Kumar. A Comparison of Document Clustering Techniques. In *KDD-2000 Workshop on Text Mining*, 2000.

[16] M. Stephens, M. Palakal, S. Mukhopadhyay, R. Raje, and J. Mostafa. Detecting gene relations from medline abstracts. In *Pacific Symposium on Biocomputing*, volume 6, pages 483–96., 2001.

[17] D. M. Wilkinson and B. A. Huberman. A method for finding communities of related genes. In *Proceedings of the National Academy of Sciences*, 2003.

[18] J. D. Wren and H. R. Garner. Shared relationship analysis: ranking set cohesion and commonalities within a literature-derived relationship network. *Bioinformatics*, 20(2):191–8., 2004.